La « mise en variables » des textes : un sujet de controverses

Gaël de PERETTI

INSEE, Département des prix à la consommation, des ressources et des conditions de vie des ménages

After language, our greatest invention is numbers. Numbers make measures and maps and so enable us to figure out where we are, what we have and how much it's worth.
(Wright, B. D., « A History of Social Science Measurement ». *Educational Measurement : Issues and Practice* (Winter 1997) 33-45, 52.)

Si l'emploi de questions ouvertes dans les questionnaires s'est développé ces dernières années, leur exploitation reste relativement rare (en particulier à l'Insee[1]) du fait de la difficulté des traitements de ce type de données. On peut distinguer trois temps : la collecte ; la « mise en variables » des réponses ; et l'analyse des réponses. Nous nous concentrerons plutôt sur le deuxième temps mais il semble nécessaire de décrire les deux autres car ils sont bien évidemment liés. L'ensemble de ce processus[2] va nécessiter la manipulation de trois savoirs complémentaires et différents : la statistique, les sciences du langage et les sciences sociales. Cette difficulté supplémentaire peut aussi être perçue comme un avantage dans le sens où elle offre un vaste champ de recherches pluridisciplinaires (d'Aubigny, 2001). Ceci peut en partie expliquer l'engouement récent pour ce type d'analyse.

Outre l'intérêt intellectuel que suscitent ces analyses, quelles sont les raisons qui peuvent expliquer cet enthousiasme ? Tout d'abord, les questions ouvertes sont particulièrement performantes lorsque l'on souhaite recueillir une information spontanée (d'Aubigny, 2001). Ensuite, des évolutions technologiques ont modifié le coût et la faisabilité du recueil de ces informations (Collecte Assistée Par Informatique ou CAPI). Enfin, l'engouement pour ce nouveau

---

[1] l'Insee a peu publié d'études s'appuyant sur des analyses textuelles si ce n'est l'étude de Christian Baudelot et Michel Gollac sur « Bonheur et travail » réalisée certes à partir d'une partie variable de l'enquête permanente sur les conditions de vie des ménages (EPCV), mais par une équipe de sociologue de l'ENS (Baudelot, Gollac, 1997). Or, les questions ouvertes peuvent être au cœur même de la problématique étudiée. Par exemple, dans la partie variable « Transmissions familiales » de l'EPCV, les deux questions suivantes étaient posées : « Qu'est-ce que vos parents vous ont transmis ou légué de plus important ? » ; « Qu'est-ce qui est pour vous le plus important de transmettre ou de léguer à vos enfants ? ». Cette dernière a été exploitée par le DEP du Ministère de la Culture associé à l'enquête. Une autre question ouverte (« Quelle profession ou quel genre de profession souhaiteriez-vous pour votre enfant ? ») a été exploitée de façon assez sommaire dans un article récent : « Les familles ouvrières face au devenir de leur enfant » toujours par un chercheur extérieur à l'Insee (Poullaouec, 2004). Enfin, Ludovic Lebart a exploité, à la demande de l'Insee, les deux questions ouvertes de l'enquête sur le devenir des allocataires du RMI mais dans une visée pédagogique ou comme il le dit lui-même : « […], explorer les possibilités et avantages de ce type de recueil ».

[2] Ce processus est valable pour toute étude statistique et correspond au passage de *data* (information collectée par le statisticien) à *given* (information offerte par le statisticien après apurement, redressement, imputation, etc.) pour reprendre ces anglicismes chers à Jean-Claude Deville.

champ d'investigations a entraîné la multiplication[3] d'outils informatisés qui non seulement ont simplifié le traitement (gain de temps en particulier) mais l'ont rendu plus efficace (en termes d'informations produites).

Nous allons nous intéresser à un point particulier de ce vaste domaine qu'est l'analyse textuelle[4] : le traitement des questions ouvertes dans les enquêtes auprès des ménages à l'aide d'une analyse lexicométrique. Pour schématiser, cette méthode se résume à la constitution d'un « tableau lexical entier » (TLE), comprenant en colonne, l'ensemble des mots utilisés dans le corpus textuel défini par l'ensemble des réponses et en ligne, les fréquences d'utilisation de chaque mot pour un enquêté. Par la suite, les analyses statistiques consistent à comparer les différents profils lexicaux.

Dans cette présentation, nous allons dans un premier temps revenir sur l'intérêt des questions ouvertes et de l'analyse textuelle. Puis, au travers de la question finale de l'enquête sur les usagers des services d'hébergement et de distribution de repas chauds dite « sans-domicile 2001 » - « Souhaitez-vous ajouter des informations que ce questionnaire n'a pas permis de recueillir ? » - nous regarderons l'impact de la collecte sur ce type de questions. Enfin, nous développerons la phase de « mise en variables » des réponses à cette question.

1. Les questions ouvertes : intérêt et traitement

Qu'est-ce que l'on mesure ?

L'introduction d'une question n'a de sens que si elle apporte une information pertinente. Ceci est valable aussi bien pour les questions fermées que pour les questions ouvertes. Ces dernières doivent faire face à trois problèmes.

Le premier a été et est encore le sujet de débats particulièrement aux Etats-Unis à propos des enquêtes d'opinion. La critique la plus fréquente sur l'utilisation de questions ouvertes dans les enquêtes concerne les difficultés d'expression de certains enquêtés qui les empêcheraient de produire une réponse claire alors qu'ils peuvent avoir une idée précise sur le sujet abordé. Ainsi, les questions ouvertes mesureraient plus leur niveau d'éducation que leur position sur le sujet (Craig, 1985). *A contrario*, Geer montre lors d'une étude que les personnes qui ne répondent pas à une question ouverte le font par manque d'intérêt[5] sur le sujet étudié plutôt que du fait d'une incapacité à répondre à ce type de question (Geer, 1988).

L'autre critique concerne la pertinence de l'information collectée. Généralement, les enquêtés seraient « peu susceptibles de sonder leur mémoire de façon assez précise pour se rappeler correctement des informations qui ont généré leur jugement global » sur la question posée (Smith, 1989, p84). Plus précisément, « les commentaires ne révèleraient pas leur opinion fondamentale. Au contraire, les réponses révèleraient des goûts ou des aversions plus superficiels, comme ceux que l'enquêté aurait pu lire récemment dans les journaux ou entendre à la télévision ou lors d'une conversation avec un ami » (Smith, 1989, p84). Plus généralement, les opposants aux questions ouvertes pensent qu'elles génèrent de simples stéréotypes. A l'opposé, les partisans des questions ouvertes considèrent qu'elles permettent aux enquêtés de définir elles-mêmes leur propre champ et de nommer les problèmes qui les concernent directement (Kelley, 1983). Parallèlement, Geer montre que : premièrement, les questions fermées sont tout aussi sensibles que les questions ouvertes à l'actualité ; deuxièmement, l'influence de l'actualité s'exerce essentiellement sur les sujets d'intérêt des enquêtés et de toutes façons, si l'information est pertinente, il semble logique qu'elle soit intégrée dans le discours des personnes (Geer, 1991). Toutefois, il est nécessaire de

[3] Jacques Jenny donne en 1997 une liste très complète des logiciels d'analyse textuelle développés et utilisés en France (Jenny, 1997)
[4] Dans l'article, cité précédemment, Jacques Jenny dresse une liste très complète des différents champs et méthodes de l'analyse textuelle (*op. cit.*)
[5] Les partisans de Craig pourraient répliquer que l'intérêt n'est pas indépendant du niveau d'éducation sous l'hypothèse que plus le niveau d'éducation est élevé plus on aurait de centres d'intérêt mais ceci reste à discuter.

garder présent à l'esprit que l'actualité aussi bien collective (événement ayant eu des répercussions médiatiques) qu'individuelle (événement qui vient d'affecter directement la vie de l'enquêté) influence la réponse de l'enquêté. L'objectif ici n'est pas de relancer le vieux débat sur la supériorité des questions ouvertes ou fermées qui existe depuis au moins soixante ans[6]. Mais si, jusqu'à présent, les questions fermées ont été privilégiées principalement car elles étaient plus faciles à poser, à coder et à analyser (Schuman and Presser, 1981), les récents progrès technologiques devraient relancer l'utilisation des questions ouvertes dès lors que l'on s'accorde sur la pertinence et la richesse des informations collectées grâce à ces dernières.

Toutefois et malgré ces progrès, le dernier problème concerne justement la phase de quantification. Nous reprenons ici une terminologie développée entre autres par Alain Desrosières : « Le premier moment est celui de la quantification proprement dite. Le verbe *quantifier* est ici employé dans un sens différent de celui du verbe *mesurer*. L'idée de mesure, inspirée des sciences de la nature, suppose implicitement que quelque chose de bien réel, déjà existant, analogue à la hauteur du Mont-Blanc, peut être mesuré, selon une métrologie réaliste. En revanche, le verbe *quantifier* implique une traduction, c'est-à-dire une action de transformation, résultant d'une série d'inscriptions, de codages, de calculs, et conduisant à une mise en nombre. Celle-ci contribue à exprimer et faire exister sous une forme numérique par mise en œuvre de procédures conventionnelles, quelque chose qui était auparavant exprimé seulement *par des mots et non par des nombres.* » (Desrosières, 2004). Et cette quantification est d'autant plus délicate que les procédures de codage et de traduction ne sont pas fixées dans le marbre ni même parfois reconnues. Le matériau à transformer est très complexe et cette transformation est explicitement liée à ses parties aval (analyse ou interprétation) et amont (collecte).

Du codage à l'analyse textuelle

Pendant longtemps, les procédures de quantification des questions ouvertes ont consisté à les « fermer », c'est-à-dire à les coder. Ce traitement a suscité de nombreux débats toujours d'actualité. Tout d'abord se pose le problème de la médiation du chiffreur (ou codeur). En effet, toute codification nécessite une interprétation par le codeur de la réponse de l'enquêté[7]. Or, des travaux ont montré le biais introduit lors du codage de questions ouvertes du fait de la distance qui existe entre ce que voulait dire l'enquêté et l'interprétation qu'en a fait le codeur (Kammeyer, Roth, 1971). Ces travaux ont consisté à comparer le codage et une description détaillée des réponses par les enquêtés. Les auteurs ont aussi insisté sur le fait que l'on ne peut postuler que les erreurs commises se compensent et qu'ainsi l'articulation générale reste la même. Ceci les a conduit à conclure sur l'extrême prudence lors de l'interprétation et ce quelles que soient les précautions prises lors du codage. Ensuite, toute codification entraîne une perte de méta-information. En effet, par essence, classer ou regrouper des individus revient à supprimer de l'information. De même, coder du texte revient à réduire un corpus textuel plus ou moins dense à un simple thème. Toute l'information lexicale contenue dans les réponses en clair comme la longueur des phrases, le vocabulaire employé, la densité syntaxique, l'utilisation de verbes modaux, l'articulation des idées est généralement perdue[8]. Ces problèmes sont encore accentués quand la réponse à coder est complexe. Faut-il se contenter de retenir le thème principal abordé dans la réponse ? Quels sont les critères qui permettent de le détecter ? Dans le cas où l'on

[6] De nombreux articles traitent du sujet (par exemple : Blair, Sudman, Bradburn, Stocking 1977; Bradburn 1983; Bradburn, Sudman, and Associates 1979; Dohrenwend 1965; Dohrenwend Richardson 1963 ; Juan 1986 ; Lazarsfeld 1944 ; Schuman, Presser 1981 ; Sheatsley 1983 ; Sudman Bradburn 1982). L'ensemble de ces recherches suggère qu'il n'existe pas de format de question supérieur à l'autre dans chaque situation et que de ce fait les deux formats ont leur place dans les enquêtes.

[7] Dans les années 1960, une critique radicale de la sociologie quantitative (en particulier la sociologie d'enquête par questionnaire développée par Lazarfeld) s'appuie sur cette même idée de différence entre les schèmes de perception des acteurs (enquêtés) et des sociologues qui montre que l'erreur d'interprétation est d'autant plus grande que l'écart social et culturel est important (Cicourel, 1964). Ces critiques conduiront à une prise en compte de l'importance du questionnement, de l'interaction entre enquêteur et enquêté mais aussi des problèmes d'agrégation de réponses identiques formulées par des personnes de milieux sociaux différents.

[8] Ce problème n'est pas entièrement résolu en travaillant sur l'ensemble du texte du fait du biais possible introduit par la saisie des enquêteurs mais aussi (voire surtout) par la quantification ou « mise en variable » des textes qui, par construction, conduit à une réduction de l'information.

déciderait de conserver plusieurs thèmes, doit-on se fixer une limite en nombre de thèmes et si oui quels critères appliquer pour fixer ce nombre ? Si l'on décide de conserver l'ensemble des thèmes doit-on tenter de les hiérarchiser ou se contenter de noter leur co-occurrence ? Enfin, il faut se poser la question des réponses rares et de leur traitement. Faut-il seulement les considérer comme du bruit ou plutôt comme une information sur une certaine catégorie de la population ?

L'ensemble de ces questions a conduit des praticiens à fixer des règles de codification afin de rendre plus robuste cette opération. Ainsi, dès le début des années 1950, un certain nombre de travaux ont été consacrés au problème de codification des réponses aux questions ouvertes. En particulier, Lazarsfeld et Barton (1955) spécifiaient quatre conditions requises pour une bonne codification : la codification doit aller du général au particulier afin de permettre une analyse plus ou moins fine selon que l'on utilise les grands thèmes ou des items plus détaillés[9] ; l'articulation logique des catégories doit reposer sur un principe unique de classification et assurer à cette dernière un caractère disjonctif et exhaustif ; elle doit s'adapter à la structure de la situation analysée ; elle doit s'adapter au cadre de référence de l'enquêté. Une fois ces principes de base adoptés, reste en suspens la question des techniques de codification. Une pratique assez courante consiste à fabriquer une grille de codage[10] *a priori* à partir de la confrontation des travaux de deux codeurs différents (Montgomery, Crittenden, 1977). Trois cas sont possibles : les catégories identifiées par les codeurs se correspondent totalement (c'est-à-dire qu'elles regroupent exactement les mêmes réponses) ; plusieurs catégories de l'un correspondent à une seule catégorie de l'autre ; pas de correspondance possible. Cette première analyse permet de dégager les catégories qui seront retenues en s'appuyant sur les quatre conditions requises pour gérer le cas « pas de correspondance possible » et de préciser les instructions pour les futurs codeurs. Ensuite, l'ensemble des réponses sont analysées par différents codeurs qui disposent tous de la même grille de lecture (catégories et instructions). Dans le cas où une même réponse serait codée différemment, la règle de la majorité est appliquée (c'est-à-dire que le code le plus souvent cité est retenu). Des règles plus pratiques de post-codage peuvent aussi être appliquées. Ainsi Xavier Marc (2001) conseille de ne pas créer de catégorie « autre réponse » concernant plus de 5 % de la population enquêtée et parallèlement de ne pas retenir de thème concernant moins de 3 % de la population. L'avantage de ce mode de chiffrement est que les analyses statistiques qui vont suivre s'appuieront sur des procédures déjà consolidées puisque cette opération de codage consiste à « fermer la question ». Cependant, l'ensemble des questions soulevées précédemment pose le problème de la légitimité d'effectuer un post-codage des réponses à une question ouverte puisque, idéalement, il nécessiterait le travail de plusieurs codeurs pour assurer une certaine robustesse des résultats et que de toutes façons il entraînerait la perte d'une quantité importante d'information[11].

Ces procédures de codage ont assez peu d'inconvénients pour des réponses simples ou stéréotypées ou peu nombreuses. Mais de nombreux éléments d'analyse sont perdus lors du post-codage - qualité de l'expression, registre du vocabulaire, syntaxe, tonalité générale de l'entretien, longueur des réponses, etc. - éléments qui sont parfois liés à l'introduction de ces questions[12]. Par exemple, dans l'enquête du CREDOC sur les « Nouveaux arrivants au RMI », l'analyse de la question ouverte sur les jugements des allocataires sur le RMI n'avait pas pour but une quantification des jugements portés déjà étudiés au travers des nombreuses questions fermées qui constituaient le questionnaire mais plutôt de « renseigner sur la manière dont ils les formulent, étant entendu qu'elle mettra ainsi en valeur les affects qu'elle contient » (Aldeghi, 1988, p148). L'autre inconvénient majeur concerne les difficultés à traiter les réponses complexes ou composites qui sont selon Ludovic Lebart « littéralement laminées par le post-codage » alors même que « c'est dans ce cas que la valeur heuristique des réponses libres est la plus grande »

---

[9] C'est le principe de toute nomenclature construite en une arborescence de plus en plus fine.

[10] Cette grille peut s'appuyer sur des réflexions reposant sur les autres questions de l'enquête en particulier si la question ouverte vient compléter une batterie de questions sur un thème précis. Elle peut aussi reposer sur un sous-échantillon de réponses.

[11] Ce dernier point doit être nuancé par le fait que si ce travail de réduction n'était pas fait, il n'y aurait de fait aucune information transmise.

[12] Toutefois, rien n'empêcherait a priori un codeur de décrire chaque réponse selon une variable décrivant la forme (qualité de l'expression) et le fond (thème) de la réponse.

(Lebart, 2001). Enfin, les réponses rares ou peu fréquentes, difficiles à analyser en première lecture sont par construction affectées à la catégorie « autre réponse », ce qui empêche généralement toute interprétation de cette dernière. Or, ces réponses rares peuvent être produites par des catégories particulières d'individus et donc présenter un certain intérêt lors de l'interprétation des résultats.

Parallèlement, le développement d'outils (ou logiciels) a facilité le traitement de ces données complexes que sont les réponses libres. Cependant, ces outils ont généralement été développés dans le cadre d'un des nombreux courants de recherche ou école de pensée de l'analyse textuelle que sont la lexicométrie, l'analyse de contenu thématique, l'analyse automatique des réseaux de mots associés, l'analyse propositionnelle et prédicative du discours, etc.[13]. Aussi, le choix d'un logiciel d'analyse de données textuelles n'est pas innocent puisqu'il sous-entend l'adoption d'un cadre théorique d'analyse de discours. En particulier, Brugidou et *al* (2000) se sont interrogés sur la complexité des paramètres qui interviennent dans le choix d'un logiciel. Dans notre cas, nous allons donner les raisons qui ont guidé notre choix dans la « mise en variable » des réponses libres en insistant en particulier sur les contraintes qui pesaient sur notre matériau brut.

2. Les problèmes liés à la collecte

De l'oral à l'écrit

L'enquête a été collectée sur papier pour des raisons pratiques évidentes et dans des conditions parfois difficiles. En particulier, lors des entretiens dans les services de distribution de repas chauds, l'enquêteur ne disposait pas toujours d'un lieu réservé ou d'une table isolée pour réaliser son interview. Indépendamment des conditions de collecte, le protocole retenu impliquait le passage d'un discours oral à une retranscription écrite de l'enquêteur. Implicitement, cela suppose que l'on fasse l'hypothèse que l'écrit sera l'image fidèle de l'oral or il est clair qu'il n'existe pas de bijection entre oral et écrit dans un tel cadre (Lallich-Boidin, 2001). En effet, l'écrit est évidemment plus précis dans certains cas : marque du pluriel et du genre ; majuscule et minuscule ; sujet des verbes. La perte de ces informations est appelée la neutralisation des oppositions : opposition entre unicité et nombre, homme et femme, etc. De façon duale, l'écrit n'est pas capable de retranscrire les distinctions de l'oral (c'est ce problème auquel nous sommes confrontés).

Ainsi, comment interpréter la phrase : « plus de place(s) dans les centres ». Dans le cas où cette phrase serait extraite d'un corpus textuel plus riche, nous utiliserons le contexte pour décider si l'enquêté voulait souligner le manque de place dans les structures collectives ou demander l'augmentation du nombre de places dans les structures existantes mais rien ne permet d'affirmer que nous pourrons choisir avec certitude. Dans le cas où elle serait le corpus entier, nous pouvons décider d'appliquer une règle syntaxique. Si l'enquêteur a mis un « s », nous considèrerons que l'enquêté réclame une augmentation du nombre de places, si l'enquêteur n'en a pas mis, que l'enquêté constate un manque de place dans les structures d'accueil. Cette règle est simple à mettre en place mais elle est, à la lecture de l'ensemble des réponses, un pis-aller. En effet, il existe un nombre très important de fautes d'orthographes dans l'ensemble des corpus textuels saisis qui légitiment la remise en cause de cette procédure. Afin d'éviter ce genre de problème, des solutions ont été proposées : utilisation de logiciels de retranscription ; enregistrement des réponses avant saisie ; formation spécifique des enquêteurs à la saisie de tel corpus[14]. Ces techniques n'étaient pas présentes lors de la collecte de l'enquête « sans-domicile 2001 ». De plus, aucune instruction spécifique à cette question n'était présente dans le manuel d'instructions aux enquêteurs. La saisie des réponses a donc été source de nombreuses disparités entre les différents enquêteurs.

[13] Jacques Jenny a effectué une présentation très complète des fondements théoriques de l'élaboration des logiciels d'assistance à l'analyse de contenu et de discours disponibles (Jenny, 1997).

[14] En particulier, il serait intéressant d'ajouter si nécessaire des inflexions pour rendre compte au mieux du discours produit.

Les réponses ont été écrites par l'enquêteur puis ont été saisies par des opérateurs. De nombreux enquêteurs ont adopté un style télégraphique sans sujet ni conjugaison et utilisé des abréviations afin de collecter l'ensemble du discours produit par l'enquêté. Cependant, les réponses saisies ne pouvaient dépasser 200 caractères, source d'un certain nombre de réponses tronquées. Dans notre cas, une méthode biaisée pour estimer le nombre de réponses tronquées a consisté à compter toutes les réponses contenant deux cents caractères. Elles sont au nombre de 123 sur les 2186[15] réponses abordant au moins un sujet (c'est-à-dire différente de RAS, non ; rien à ajouter, etc.) soit 5,6 % des réponses. A priori, on peut penser qu'elle tend à surestimer le nombre de réponses tronquées puisque le discours saisi peut très bien s'arrêter effectivement à 200 caractères. On peut, parallèlement, compter les réponses de 200 caractères finissant par un mot tronqué (*« […] Il faut améliorer les conditions de v… »*) ou une phrase incomplète (*« […] souhaite avant tout pouvoir… »*). Il y a 20 phrases qui sont manifestement incomplètes et 46 phrases qui finissent par un mot tronqué. Dans ce dernier cas de figure, deux possibilités sont envisageables : soit le discours finissait par ce mot tronqué (soit par manque de place soit parce que l'enquêteur utilisait des abréviations ce qui est le cas dans un nombre important des réponses) soit le discours continuait. La lecture de certains questionnaires[16] confirme les deux cas envisagés. Ainsi, un questionnaire repéré comme tronqué finissait effectivement par le dernier mot saisi. En revanche, un autre se prolongeait longuement, la personne enquêtée donnant son opinion ou son sentiment sur de nombreux sujets : « *Difficultés particulières par rapport au logement différentes à Toulouse. C'était plus facile l'insertion à Toulouse que sur Paris. Il y a l'air d'avoir plus de structures mais moins de places. On peut tourner de foyers en foyers sans d'autres issues que par le travail car il faut la caution. Quand on est dans un foyer, c'est difficile de trouver du travail (horaires, peu de calme). Différent à Toulouse où on m'avait donné un studio sans caution. Le Samu Social (115), injoignables, j'ai bataillé et on m'a dit qu'il fallait redescendre à Toulouse, trois semaines dans la rue : « Vous n'avez qu'à redescendre en stop ». Le foyer Emmaüs : nickel au niveau propreté, le seul problème, il faut partir de 8h à 18h, toute la journée dans la rue, sans ressource et sans savoir quoi faire. Paris demande beaucoup de courage et de santé. On se sent perdu sans repère. Les domiciliations sont difficiles, les prix des transports aberrants sur un RMI. Si je n'ai pas de place dans un hôtel social, je redescendrais sur Toulouse* ».

Au-delà de ce problème de troncature, les différences liées aux prises de note (phrases complètes ou abréges, discours indirect ou direct) nous ont conduit à préférer une approche thématique à partir d'une analyse lexicométrique du fait des fortes disparités entre les corpus textuels difficilement allouables à l'enquêté et surtout à une grande prudence quant à la généralisation des résultats du fait des possibles traductions des réponses libres des enquêtés par les enquêteurs. Mais ce type d'intervention de l'enquêteur sur le corpus textuel étudié n'est pas la seule.

L'effet enquêteur

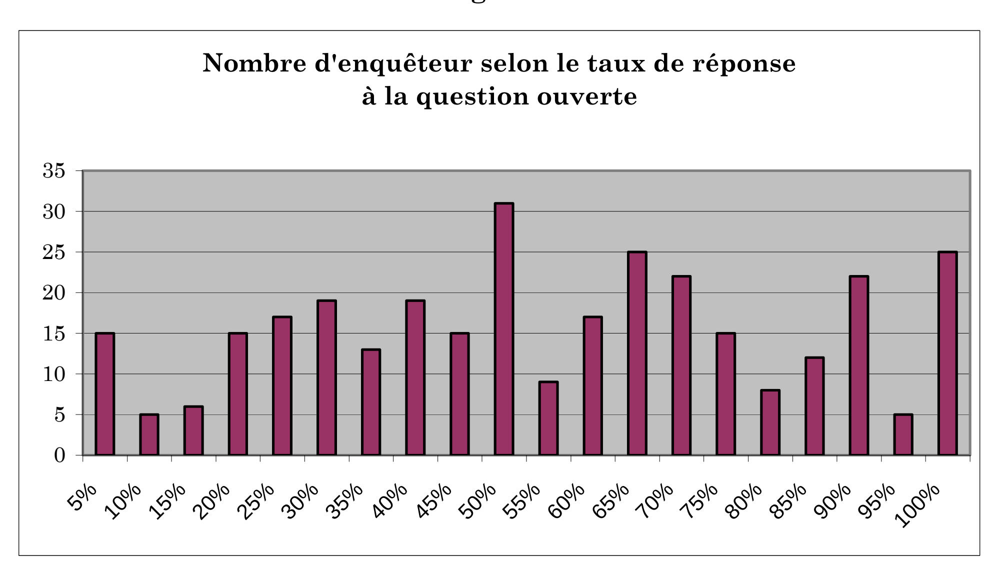
Cette question ne peut être mise de côté dès lors que l'on traite d'une étude sur une population en difficulté d'insertion sociale. En effet, lors d'une étude méthodologique sur les principes à adopter pour améliorer la qualité des enquêtes sur ce type de population (Dubéchot Legros, 1993), les auteurs montrent l'importance de contrôler au mieux l'effet enquêteur. En fonction de son âge, de son expérience, de sa formation professionnelle, la relation qu'il va établir avec l'enquêté sera de nature différente, et peut, en raison de l'impossibilité d'une standardisation parfaite de cette relation malgré les instructions données aux enquêteur, conduire à un déroulement différencié des entretiens. Cet effet est d'autant plus sensible que dans notre cas, nous nous intéressons à la question finale et qu'aucune instruction spécifique n'était donnée aux enquêteurs dans leur manuel d'instructions. Après une très longue série de questions plus ou moins personnelles voire stigmatisantes (ou perçues comme telles), le fait de répondre à cette dernière question peut être influencé par la relation qui s'est instaurée au cours de l'entretien entre enquêté et enquêteur ce qui dépend *a priori* fortement des caractéristiques propres de ce dernier.

[15] Ceci correspond à un taux de réponse de 52 % ce qui est proche du taux de réponse (59 %) à la question équivalente dans l'enquête devenir des allocataires du RMI.
[16] Ces questionnaires ont été selectionnés au hasard parmi les questionnaires en provenance d'Ile de France.

L'étude des questions ouvertes dans des enquêtes d'opinion a mis en évidence un effet enquêteur sur les réponses des enquêtés. Cependant, cet effet est généralement considéré comme mineur sur les questions d'opinion ouvertes (Caillot Moine, 2001). L'objectif de l'analyse textuelle, au-delà de l'analyse de l'univers lexical des réponses, est de croiser cet univers lexical avec les caractéristiques des individus afin de déterminer l'influence de ces derniers sur le contenu des réponses. La conclusion de l'étude de Caillot et Moine est que, si l'on constate des effets sur la forme du corpus textuel et le nombre de thèmes abordés, cette interaction ne modifie pas significativement l'information apportée par cette question et les liaisons entre les thèmes abordés et les caractéristiques des enquêtés. Du fait de la particularité de notre enquête et de la différence entre une question d'opinion classique et notre question ouverte, il est nécessaire de vérifier si dans notre cas, l'effet enquêteur est différent.

Une première approche de cet effet enquêteur consiste à regarder les taux de réponse à cette question par enquêteur et plus particulièrement la distribution de ces taux. 315 enquêteurs ont mené cette enquête et ont collecté en moyenne 13 questionnaires. Les différences entre enquêteurs sont grandes. Neuf enquêteurs ont collecté 1 questionnaire et un en a collecté 29. De même, les disparités sur les taux de réponse à la dernière question sont très fortes (voir Figure 2). Ainsi, quatorze enquêteurs ont un taux de réponse nul (ils ont en moyenne enquêté 5,6 personnes) et vingt-cinq enquêteurs ont un taux de réponse de 100% (ils ont en moyenne enquêté 8,7 personnes). En moyenne (non pondérée), un enquêteur a obtenu un taux de réponse de 53,3 % à cette question[17].

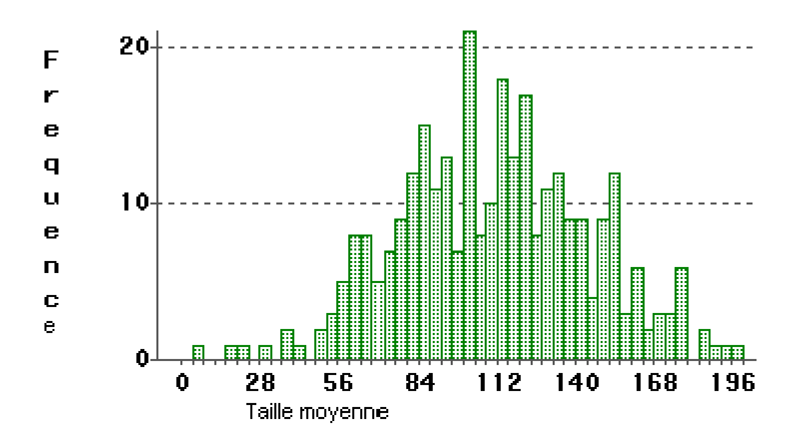
Figure 1



La deuxième approche revient à vérifier l'hypothèse de Caillot et Moine sur la longueur des textes. Nous avons donc analysé la distribution de la longueur des textes bruts. Ainsi, pour chaque enquêteur, nous avons calculé à partir des réponses non corrigées la taille moyenne des réponses en nombre de caractères. La longueur des réponses varie beaucoup (de 5 caractères à 200) et un sixième des réponses compte au moins 190 caractères. Le fait d'étudier la taille par enquêteur permet de compléter cette première étude. Une nouvelle fois, nous constatons une grande disparité dans la longueur des textes.

[17] Ce taux est obtenu sur les réponses différentes de RAS et ses déclinaisons.

Figure 2 : nombre d'enquêteurs en fonction de la longueur des réponses



Tout d'abord, l'effet lié à la troncature des réponses est de nouveau visible, la queue droite de distribution est épaisse. Ensuite, le nombre moyen de caractères pour un enquêteur donné est de 111 caractères (moyenne des moyennes par enquêteur) contre 117 sur l'ensemble des réponses. Cette différence est significative en terme statistique (test non paramétrique du fait des caractéristiques de notre variable) mais apporte a priori peu d'informations. En effet, six caractères permettent peut-être d'écrire un mot mais pas une idée nouvelle sauf dans le cas des réponses télégraphiques. Cela implique seulement que les réponses longues sont plutôt concentrées sur certains enquêteurs.

Ces approches sommaires permettent de mettre à jour les disparités de résultats de collecte entre les enquêteurs mais n'apportent pas de solution pour les corriger. Ces disparités confirment la nécessaire prudence lors de la phase d'interprétation et vont conditionner la « mise en variable » des textes comme nous allons le voir par la suite.

3. Normalisation, quasi-lemmatisation, des outils de l'analyse lexicométrique

Comme nous l'avons déjà fait remarquer, l'engouement pour l'analyse statistique de texte a entraîné le développement parallèle d'un grand nombre de logiciels ayant chacun ses spécificités propres dépendant complètement de l'approche du concepteur. Or, le choix (plus ou moins contraint) d'un logiciel a des conséquences sur le type d'analyse que l'on peut envisager. Nous avons travaillé sur le logiciel SPAD-T dont la philosophie générale consiste à repérer l'ensemble des mots (formes graphiques) utilisés dans le corpus textuel et de calculer leur occurrence. Dans ce logiciel, avant tout traitement statistique, il est nécessaire d'effectuer deux procédures (ou « méthodes » selon la terminologie propre au logiciel) qui permettent respectivement de repérer les mots utilisés mais aussi les groupes de mots (ou « segments répétés »). Parallèlement, il est possible d'effectuer des corrections et des regroupements de mots ou segments afin d'en réduire le nombre tout en limitant la perte d'information et d'obtenir ce que nous appellerons notre vocabulaire d'étude. L'ensemble des traitements statistiques développés par le logiciel repose sur une analyse lexicométrique du corpus que l'on peut développer selon deux axes : le poids absolu d'une forme graphique dans l'ensemble du corpus étudié ou le poids relatif d'une forme graphique au sein de différentes catégories de population. Enfin, il est possible de créer un tableau contenant en ligne des individus avec leurs caractéristiques (âge, sexe, etc.) mais aussi la fréquence d'utilisation des mots retenus dans le vocabulaire d'étude afin de caractériser la réponse à la question ouverte des individus. Ce tableau (le TLE) permet de réaliser l'ensemble des traitements statistiques connus (classification, analyse factorielle, régression), chaque mot du vocabulaire étant devenu une variable.

Cette approche peut paraître paradoxale, car comme le rappelle Dominique Labbé (2001) : « *Peu de mots dépassent le seuil de 1 % de fréquence relative et ce ne sont probablement pas les plus intéressants puisque, selon le vieil adage classique, la quantité d'information véhiculée par un mot est inversement proportionnelle à sa fréquence d'apparition* ». Ceci est confirmé par la liste des vingt mots les plus fréquemment utilisés dans les réponses à notre question ouverte et qui représente près d'un tiers des formes utilisées (32,8 %) : *de, je, un, les, pas, à, pour, et, est, la, le, d, des, que, l, en, j, on, ne, il*. L'objectif premier est donc de réduire le nombre de mots (formes graphiques) que nous prendrons en compte dans les analyses statistiques du fait des particularités de notre corpus (liées à la collecte) tout en limitant la perte d'information par rapport à notre axe d'interprétation des résultats. Cette réduction a pour but de faciliter et rendre plus robuste les calculs qui seront menés par la suite afin d'éviter de travailler sur des tables immenses et pleines de zéro. En revanche, il est nécessaire de bien définir les traitements que l'on va opérer car ils auront de fait des conséquences sur les résultats produits par les calculs. Ces effets devront donc être pris en compte lors de l'interprétation des résultats. Parallèlement, les règles de traitements que l'on définit doivent être en accord avec les visées finales de l'étude. Dans notre cas, les contraintes imposées par la collecte nous ont conduit à viser une analyse de contenu de type thématique. C'est dans cette phase que l'on retrouve les tensions entre objectivité et subjectivité. Nous allons fixer un certain nombre de conventions qui vont modifier l'objet sur lequel nous travaillons. Parallèlement, ces conventions doivent apporter de la robustesse (au sens statistique) aux procédures d'analyse statistique qui seront produites par la suite. Il est nécessaire de préciser ces conventions mais aussi de tenter de cerner du mieux possible les conséquences qu'elles auront sur les analyses qui vont suivre.

La normalisation

Avant de se lancer dans une analyse textuelle des réponses, nous avons pris le parti d'effectuer des corrections sur le texte saisi. L'objectif de cette normalisation des réponses est de « débruiter » au maximum les réponses du fait des fortes disparités dans les procédures de recueil des réponses des enquêtés. Le problème majeur de la correction est qu'elle implique nécessairement une interprétation de la part du correcteur (Lallich-Boidin, 2001). Ceci conduit à appliquer un traitement standard à l'ensemble des réponses et à fixer des normes liées aux corpus et aux traitements visés.

La matière brute sur laquelle nous avons travaillé est un texte en lettres capitales, transcription intégrale (dès lors que la réponse contient moins de 200 caractères) du texte écrit par l'enquêteur. L'ensemble des textes a été retranscrit en minuscules accentuées afin d'éviter les confusions du type « a » et « à », « bornes » et « bornés ».Tous les noms propres commencent par une majuscule et les mots composant le nom d'une association sont collés. Ainsi les « Restos du cœur » deviennent les « Restosducoeur ». Toutes les dates sont écrites sous la forme « jour mois année » avec *jour* et *année* en chiffre. Toutes les sommes d'argent sont sous la forme xxxf avec xxx correspondant au montant évoqué et f à franc (enquête réalisée en 2001). Nous avons systématisé le recours aux acronymes ou aux sigles avec une orthographe unique en lettres capitales. Enfin, nous avons transformé l'ensemble des textes en les mettant au discours direct.

Toutefois, nous avons conservé l'organisation grammaticale de la réponse. Ainsi la réponse « un travail et un logement » n'est pas modifiée alors que la réponse « il voudrait un logement et un travail » devient « je voudrais un logement et un travail ». Cette correction est un choix pratique qui n'est pas entièrement satisfaisant. Il est en effet impossible de savoir dans les deux cas précédents si l'enquêteur a réellement transcrit le discours de l'enquêté. Les deux personnes ont très bien pu dire : « *j'aimerais bien avoir un logement et un travail* ». Les deux transcriptions précédentes sont possibles car elles conservent les thèmes évoqués. Nous avons choisi de modifier le moins possible le corpus de base. Ceci a pour conséquence que dans la suite de l'analyse, les trois réponses suivantes seront équivalentes : *un travail et un logement ; avoir un travail et un logement ; j'aimerais bien avoir un travail et un logement*. En revanche, la réponse « *trouver un travail et un logement* » est différente puisqu'elle implique plus directement l'enquêté. Dans les cas précédents, il s'agit seulement d'un souhait, alors que dans le dernier exemple, l'utilisation du

verbe trouver laisse supposer qu'il y aura une recherche de sa part. Cette distinction est nécessaire dès lors que l'on s'intéressera non pas seulement aux thèmes abordés mais aussi à la formulation et à la constitution de « posture » (Reinert, 2001), c'est-à-dire de discours types relatifs à certaines catégories de personnes.

D'autres travaux auraient pu être envisagés, en particulier, le traitement de la polysémie et de l'homographie assez fréquentes dans la langue française. Dans les deux cas, l'objectif est d'ajouter des marqueurs qui permettent de repérer les différents sens d'un même mot afin d'éviter des problèmes d'interprétation des résultats.

Dans le deuxième cas, des règles syntaxiques permettent de séparer les homographes en associant à chaque forme une catégorie grammaticale. Ainsi, des chercheurs ont développé des logiciels contenant des nomenclatures de mots français, en regroupant leurs différentes flexions sous un même « lemme » étiquetté de sa forme grammaticale. Par exemple, toutes les formes conjuguées d'un même verbe sont réunies sous le doublon [nom du verbe à l'infinitif, verbe]. De même, toutes les déclinaisons de l'article « le » (la, l', les) sont réunies sous le doublon [le, article]. Ensuite, l'ensemble des règles syntaxiques de la langue française sont connues du logiciel ce qui lui permet à la lecture du texte de séparer l'essentiel des ambiguïtés. Cela permet de distinguer automatiquement les deux sens du mot « être » dans les phrases suivantes : « *[…] Faut connaître les difficultés pour un être humain de vivre dans la rue[…]* » où *être* est un nom masculin et « *[…] N'ai pas l'impression d'être aidé […]* » où être est un verbe.

La polysémie est plus délicate à traiter puisque le sens va dépendre du contexte. Ainsi, le verbe « sortir » a de multiples sens dans la langue française. Le sens le plus fréquemment retrouvé dans les réponses étudiées est celui de quitter la situation de précarité actuelle, de s'en sortir : « *[..] C'est difficile de s'en sortir avec des dettes[…]* ». Ensuite, sortir est utilisé dans les sens d'aller hors d'un lieu : « *[..] Où est ce que l'on va quand on doit sortir du foyer à 8h du matin [..]*». Enfin, il est utilisé dans le sens d'aller hors de chez soi pour aller se distraire : « *[..] Je souhaiterais pouvoir sortir plus le soir en semaine [..]* ». Pour le premier cas, le logiciel SPAD, en repérant le segment répété « en sortir », permet d'éviter la confusion de sens. En revanche, il n'est possible de distinguer les deux autres cas qu'à la lecture de la réponse. C'est évidemment une limite de l'approche lexicométrique. Toutefois, comme nous allons le voir par la suite, il est possible de prendre en compte le contexte dans certains cas pour éviter cette confusion.

La lemmatisation

Dans un deuxième temps, nous avons travaillé sur la lemmatisation de notre corpus, c'est-à-dire à donner à un mot du discours une forme canonique servant d'entrée de dictionnaire. Cette procédure correspond tout à fait au double objectif de réduction du nombre de mots et de limitation de la perte d'information. L'idée est de regrouper sous un même *lemme*, différents mots dont le sens est identique afin de lui donner plus de poids mais surtout afin d'éviter de ne pas les prendre en compte du fait de la disparité des formes utilisées. En effet, nous avons fait le choix de supprimer de l'étude tous les lemmes n'apparaissant pas au moins 15 fois dans l'ensemble du corpus.

La procédure retenue est simple. Nous avons regroupé toutes les conjugaisons d'un même verbe sous la forme infinitive si elle existait sauf si une forme était largement majoritaire (fréquence trois fois supérieure aux autres formes). Dans ce cas, le lemme correspondait à cette dernière forme. Les formes au pluriel et au singulier sont regroupées sous un même lemme sauf si leur utilisation correspond à deux sens différents et que leurs fréquences respectives permettent de les conserver toutes les deux. Ainsi les mots « personne » et « personnes » correspondent à deux lemmes différents. Le mot au singulier correspond généralement a un usage négatif : « *Après avoir vécu 20 ans dans mon pays, personne ne veut m'aider c'est lamentable* ». Le mot au singulier vise à désigner des individus avec lesquels l'enquêté est entré en contact ou une catégorie de gens : « *Dans les associations, certaines personnes ne nous aident pas* » ; « *Que les organismes type HLM puissent accepter des personnes au RMI* ». Afin d'éviter la perte de certains thèmes, nous avons eu dans certains cas une approche quasi thématique en nous appuyant sur le contexte des

réponses. En effet, il est possible de connaître pour tous les mots de notre vocabulaire, les phrases dans lesquelles ils sont utilisés. Cette contextualisation assure une certaine robustesse à la lemmatisation. Ainsi, le lemme « conjoint » regroupe les mots ou segment suivants : concubin, concubine, compagne, compagnon, copain, copine, épouse, époux, mari, ma femme. Nous avons dû retenir seulement le segment « ma femme », car le mot « femme » recouvrait d'autres sens que celui d'épouse.

Enfin, afin de définir complètement le « vocabulaire » sur lequel nous avons travaillé, nous avons supprimé tous les mots outils. En effet, ces derniers sont très fréquemment employés et apportent peu d'information sur le contenu du texte. Ces derniers n'auraient d'intérêt que s'ils n'étaient pas aléatoirement répartis parmi les différentes catégories étudiées. Or, dans notre cas, cette répartition non aléatoire pourrait s'expliquer en (grande) partie par la disparité des techniques de recueil des réponses des enquêteurs. Nous avons donc préféré les supprimer. L'opération de lemmatisation incluant aussi des regroupements thématiques et de suppression des mots outils s'appelle une quasi-lemmatisation (Lebart, 2000).

Construction des données et interprétation

Les réponses complètes donnent lieu à un corpus de 47879 occurrences (longueur totale de l'ensemble des réponses en nombre de formes graphiques) pour les 2186 réponses différentes de RAS (et toutes les variantes). Les répondants ont utilisé 4588 mots distincts (formes graphiques) soit 9,6 % des occurrences[18]. Parmi ces mots différents, 52 % sont des hapax (forme graphique n'apparaissant qu'une fois). Les procédures de normalisation et quasi-lemmatisation ont réduit le vocabulaire étudié à 438 lemmes (ce sont soit des mots, soit des lemmes, soit des segments répétés). L'ensemble des procédures statistiques que nous allons commenter s'appuie sur ce vocabulaire réduit (ou sur une extraction de celui-ci). Ce vocabulaire représente à lui seul 57,4 % des formes graphiques, sachant que les mots outils supprimés (articles, certaines prépositions ou pronoms relatifs) représentent 34,9 % des formes graphiques. Les choix retenus pour la constitution du vocabulaire ont donc fortement réduit le nombre de mots (moins de 10 % de l'effectif de départ) tout en conservant une grande partie du contenu thématique (au pire 75 %)[19].

L'objectif de cette réduction est de travailler sur des tableaux lexicaux plus petits ce qui assure une meilleure robustesse des calculs. Parallèlement, dans une optique de classification, elle peut être à l'origine de la suppression de certaines classes de petits effectifs produisant un discours très particulier. Ces choix sont évidemment discutables puisqu'ils introduisent une part de subjectivité dans un processus de quantification qui se veut au départ une méthode d'objectivation des résultats. Mais cette subjectivité est considérée comme une réponse possible aux problèmes posés par les données brutes et les résultats que l'on compte en tirer : elle doit être reconnue et assumée. En effet, notre hypothèse est que ces réponses sont autant de vues différentes sur le monde des usagers des services d'aide à travers le prisme de l'enquête « sans-domicile 2001 » et que parmi ces vues partielles de ce monde, certaines renvoient à une même chose, une même idée.

L'approche quantitative des textes comme méthode pertinente s'appuie sur trois hypothèses que nous allons décrire en reprenant la métaphore de la ville développée par Saadi Lahlou (1995). Chaque réponse est analogue à une photographie prise par un touriste à Paris. A partir de ces centaines de clichés, on peut tenter de reconstituer les monuments de Paris. Ainsi à partir des similarités entre les réponses, elles seront regroupées et assimilées à un objet ou une idée qu'elles seraient censées décrire. Comme la reconstitution de Paris sera d'autant plus pertinente que l'on connaît bien cette ville, l'interprétation des résultats et des classes produites sera d'autant plus pertinente que l'objet analysé est connu par l'analyste. Ceci pour rappeler que, quel que soit le degré d'objectivité des objets manipulés[20] et des techniques utilisées[21], la phase d'interprétation

---

[18] A titre de comparaison, dans l'enquête devenir des allocataires du RMI, les 2 010 réponses contenaient 40 004 occurrences et 4 003 formes graphiques soit 10 % des occurrences.

[19] Cette valeur est obtenue en rapportant la part de formes graphiques que constitue notre vocabulaire d'étude au vocabulaire initial privé des mots outils, ces derniers ne pouvant être vecteurs de thème.

[20] Ceci fait référence à l'échelle hiérarchisée des mesures en fonction de leur éloignement de la métrique *stricto sensu*.

fera nécessairement appel à la connaissance et aux rapports entretenus avec le sujet par l'analyste ce qui peut être source d'une certaine subjectivité. Cette dernière phase devrait faire l'objet d'une publication dans un numéro spécial d'une revue consacré à l'enquête « sans-domicile 2001 ».

Bibliographie

[1] Aldeghi I., « Etude complémentaire sur les nouveaux arrivants au RMI : apports du RMI, évolution de la situation matérielle, opinions sur le dispositif », CREDOC, *Collection des rapports*, n° 196, octobre 1998.

[2] d'Aubigny G., « Introduction », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 1-5, 2001.

[3] Baudelot C., Gollac M., « Faut-il travailler pour être heureux », *Insee-première*, n° 560, décembre 1997.

[4] Blair E., S. Sudman, N. M. Bradburn and C. B. Stocking, « How to ask questions about drinking and sex : response effects in measuring consumer behavior », *Journal of Marketing Research*, 14, pp 316-321, 1977.

[5] Bradburn, N.M., « Response Effects », in Handbook of Survey Research, P.H. Rossi, J.D. Wright and A. B. Anderson , eds. New York : Academic Press, 1983.

[6] Bradburn, N.M., S. Sudman, « Asking Questions », San Francisco : JosseyBass, 1982.

[7] Bradburn, N.M., S. Sudman, and Associates, « Improving interview method and questionnaire design : response effects to threatening questions in survey research », San Francisco : Jossey-Bass, 1979.

[8] Brugidou M., Escoffier C., Folch H., Lahlou S., Le Roux D., Morin-Andreani P., Piat G., « Les facteurs de choix et d'utilisation de logiciels d'analyse de données textuelles », Actes des Journées internationales d'Analyse des Données Textuelles, 2000.

[9] Caillot P., Moine M., « Mais quelle est la réponse ? », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 73-90, 2001.

[10] Cicourel A., « Method and Measurement in Sociology », The Free Press of Glencoe, New York, 1964.

[11] Craig S. C., « The decline of partisanship in the United States : a reexamination of the neutrality hypothesis », *Political Behavior*, n° 7, pp 57-78, 1985.

[12] Desrosières A., « Pour une politique des outils du savoir : le cas de la statistique », Communication à la conférence « Politics and knowledge : Democratizing knowledge in times of the expert », Université de Bergen, 21-22 juin 2004.

[13] Dohrenwend, B.S., « Some effects of open and closed questions on respondents' answers », *Human Organization*, 24, pp 175-184, 1965.

[14] Dohrenwend, B.S., S. A. Richardson, « Directiveness and non-directiveness in research interviewing : a reformulation of the problem », *Psychological Bulletin*, 60, pp 475-485, 1963.

[15] Dubéchot P., Legros M., « La qualité des enquêtes auprès de populations en difficulté d'insertion sociale », CREDOC, *Cahier des recherches*, n° C47, 1993.

[21] De même, que le statisticien préfère manipuler des nombres que des classes, il accorde généralement plus de crédit aux régressions qu'aux analyses de données à la française du fait de la possibilité de parler des effets d'une variable « toutes choses (in)égales par ailleurs ».

[16] Geer J. G. (1988), « What do open-ended questions measure », *Public Opinion Quaterly*, vol 52, pp 365-371.

[17] Geer J. G. (1991), « Do open-ended questions measure salient issues ? », *Public Opinion Quarterly*, vol 55 (3), pp 360-370.

[18] Jenny J, (1997), « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine ; états des lieux et classification », *Bulletin de Méthodologie Sociologique*, n°54, mars 1997, pp. 64-112

[19] Kelley S. (1983), « Interpreting elections », Princeton, Princeton University Press.

[20] Labbé D. (2001), « Normalisation et lemmatisation d'une question ouverte : les femmes face au changement familial », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 37-57.

[21] Lahlou S. (1995), « Vers une théorie de l'interprétation en analyse statistique des données textuelles », JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data. Bolasco S., Lebart L., Salem A. (eds). CISU, Roma, 1995, Vol I, pp 221-228.

[22] Lallich-Boidin G. (2001), « Données linguistiques et traitement des questions ouvertes », *Journal de la Société Française de Statistique*, tome 142, vol n°4, pp 29-36.

[23] Lazarsfeld, P.E. (1944), « The controversy over detailed Interviews -- an offer for negotiation », *Public Opinion Quarterly*, 8, pp 38-60.

[24] Lazarfeld P. F., Barton A. H. (1955), « Some general principles of questionnaire classification », in Paul F. Lazarfeld and Morris Rosenberg (eds), The language of social research, Glencoe Ill. : The Free Press.

[25] Lebart L., Salem A. (1994), « Statistique textuelle », Paris, Dunod, 1994.

[26] Lebart L. (2001), « Traitement statistique des questions ouvertes : quelques pistes de recherche », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 7-20.

[27] Lebart L. (2000), « Traitement statistique des questions ouvertes de l'enquête sur le devenir des personnes sorties du RMI ».

[28] Marc X. (2001), « Les modalités de recueil des réponses libres », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 21-28.

[29] Montgomery A. C., Crittenden K. S. (1977), « Improving coding reliability for open-ended questions », *Public Opinion Quaterly*, vol 41, pp 235-243.

[30] Poullaouec T. (2004), « Les familles ouvrières face au devenir de leurs enfants », *Economie et statistique*, n°371, décembre 2004.

[31] Reinert M. (2001), « Approche statistique et problème du sens dans une enquête ouverte », *Journal de la Société Française de Statistique*, tome 142, vol 4, pp 59-71.

[32] Schuman H, Presser S. (1981), « Questions and answers in attitude surveys », New York : Academic Press.

[33] Sheatsley P.B. (1983), « Questionnaire construction and item writing », in Handbook of Survey Research, P.H. Rossi, J.D. Wright and A.B. Anderson , eds. New York : Academic Press.

[34] Smith E. R. A. N. (1989), « The unchanging american voter, Berkeley », University of California Press.